vCluster: Disaggregation of Data Center Resources

Whitepaper | CALIENT Solution Brief

Data centers have long focused on virtualization of resources and convergence to IP for interconnectivity. This allowed the industry to scale to the mega data centers we see today. To continue to meet demand, data centers must now shift focus to the disaggregation of resources. The components of a compute node, processor, memory, storage and accelerators, need to be pulled apart into separate racks to allow for:

- more efficient compute resource utilization
- optimization of equipment racks for space, power, and cooling.
- Moreover, disaggregation means the data center operators can separate hardware lifecycles by component and as well as allow even greater resource commoditization.

Disaggregation of Compute Resources

- Separate the core components of compute nodes: compute, memory, accelerators, storage, etc.
- Compute clusters made up of specialized equipment racks will become the core building blocks of the data center. The rack is no longer the building block of the data center.
- The compute node interconnect physical layer will be optical
- The interconnect protocol between compute resources within a compute node must be native. TCP/IP is not viable for compute node interconnects.
- Composition of compute nodes cannot be limited by hardware defined infrastructure (HDI) boundaries

Physical separation of compute resources necessarily requires a physical bus capable of reaching across rack boundaries. The recent increase of serial data rates of optical interfaces to 100Gb/s and beyond provides the physical layer solution. Furthermore, the dramatic reduction in cost of these devices made possible by silicon photonics makes optical interfaces economically viable.

To realize the benefit of resource pooling, changes in resource consumption within the compute node, and even responding to resources failures, a switching fabric across the compute cluster is required. A common protocol, such as TCP/IP or SAN, with an electrical switching fabric doesn't come close to addressing the need of data bus interconnects. Even with the fastest switches, implementation of the protocol stack alone is way outside the bounds of meeting the requirements of compute node interconnects.

Compute Node Interconnect Requirements

- This is not a networking problem; it is a compute node architecture problem. Node interconnects most replicate data buses, not networking links.
- Latency: this goes well beyond ultrafast networking. Latency needs to be speed of light; even physical distance inside the data center will limit the size of a compute cluster.
- Protocol Agnostic: the interconnect protocol must be as native as possible. The interconnect protocol will depend on the specific device, and no single protocol will meet the disparate requirements of all resource types within a compute node.
- Throughput: the interconnect must continuously and deterministically operate at the full native bus rate.

The technical constraints on the physical layer interconnect eliminate the use of an electrical fabric.

- Electrical fabrics require the use of common protocols. Even when operating as a circuit switch, the intermediate fabric must support at least a framing protocol to establish a link with the client device.
- Electrical fabrics have buffer delays. Even the fastest cut-through switches have silicon I/O.
- Electrical fabrics must have an OEO interface. The OEO alone introduces serialization delay. For parallel to serial I/O such as 40G and 100G, this delay can be significant
- Electrical fabrics have cost. The cost is largely mitigated by the gain in resource utilization of the compute resources, but when compared to Calient's all-optical switch fabric, it is cost that is completely unnecessary.

Calient's optical circuit switch (OCS) meets all of the requirements as a switching fabric inside the compute node. Composition of a compute node from pooled resources inside a compute cluster requires the establishment of point-to-point links.

- The OCS is completely protocol agnostic. An OCS fabric will support any native protocol and at any data rate.
- The OCS has speed of light latency. There is no OEO conversion and no buffering.
- To maximize the economic benefit of resource pooling, the OCS must be large enough to support all of the devices in a compute cluster
- The size of the compute cluster is defined by the speed of light latency of the data bus.

- Calient's S-Series and Edge Series fabrics are ideally sized for the compute cluster.
    - The OCS is very low cost.
- With no intermediate active optics devices and 45W total system power, the OCS adds very little total cost to the compute cluster.
- Because the OCS is protocol and data rate agnostic, it will have the longest lifetime of any device in the compute cluster.

Calient's OCS provides a technically and economically viable solution to creating virtualized compute clusters. The OCS fabric meets all of the requirements for switching compute node interconnects and at a faction of the cost of any electrical fabric.

CALIENT Core Technology

CALIENT's Optical Circuit Switch is a large port count all-optical (OOO) switch that establishes, monitors and switches physical layer connections between single-mode optical fibers using Micro Electro Mechanical Systems (MEMS) based optical switching. Connections are made between fibers carrying signals with any data rate or protocol. Any input fiber on the S-Series OCS can be connected to any output fiber making a fully non-blocking switch fabric.

Light is directed from the input fibers to the output fibers using arrays of tiny silicon mirrors that are fabricated using the proven CALIENT MEMS process. An optical signal transmitted through the OCS passes through three sections of the switch core: the input collimator array, which directs the light from each input fiber to its input mirror; the mirror matrix, an array of MEMS input mirrors and an array of MEMS output mirrors; and the output collimator array, which couples light from each output mirror back into its output fiber. High-quality mirrors and collimators and precise electrostatic control of the position of each mirror, enable typical switch times of less than 50ms and optical loss that is less than 3.0 dB for CALIENTs complete line of optical circuit switches.

The CALIENT S320 and Edge|640 Optimization Solution

CALIENT Technologies using the S320 and Edge|640 Optical Circuit Switches has solved this challenge by providing flexible optical layer on-off ramps between different equipment and network domains. As shown in Figure 1, adding Optical Circuit Switches (OCS) at the edges of metro and wide area networks allows the multi-layer control plane to access and select resources from any domain or vendor, including a legacy network.

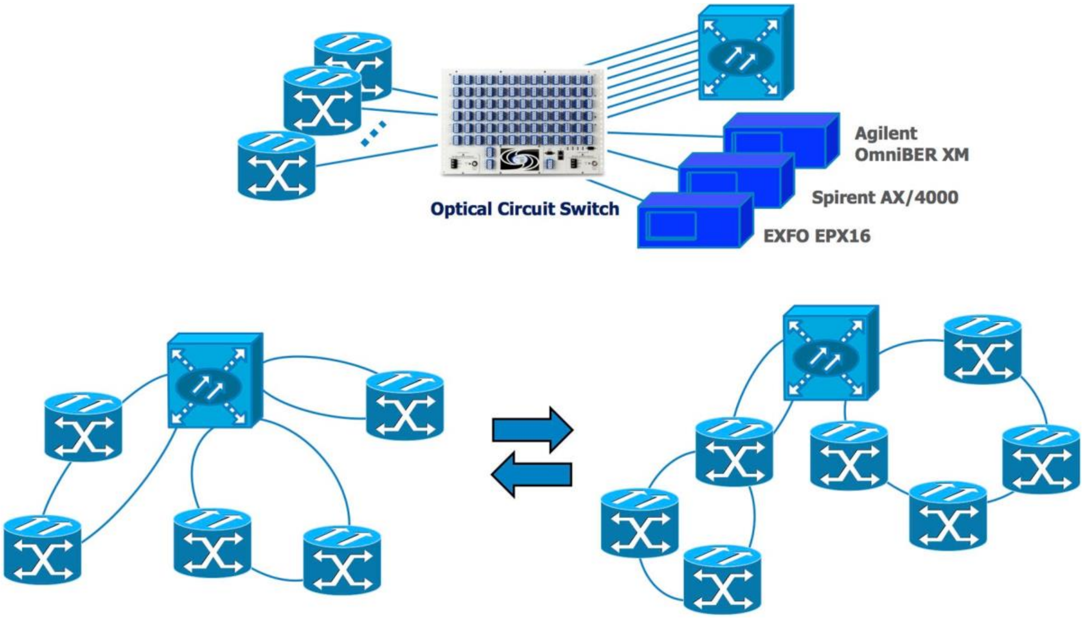

Figure 1: Improving Test Lab Facility ROI with CALIENT Optical Topology Reconfiguration.

CALIENT’s OCS technology delivers this capability in an extremely reliable, field proven, low energy cost effective package.

CALIENT Core Technology

CALIENT’s Optical Circuit Switch is a large port count all-optical (OOO) switch that establishes, monitors and switches physical layer connections between single-mode optical fibers using Micro Electromechanical Systems (MEMS) based optical switching. Connections are made between fibers carrying signals with any data rate or protocol. Any input fiber on the S-Series OCS can be connected to any output fiber making a fully non-blocking switch fabric.

Light is directed from the input fibers to the output fibers using arrays of tiny silicon mirrors that are fabricated using the proven CALIENT MEMS process. An optical signal transmitted through the OCS passes through three sections of the switch core: the input collimator array, which directs the light from each input fiber to its input mirror; the mirror matrix, an array of MEMS input mirrors and an array of MEMS output mirrors; and the output collimator array, which couples light from each output mirror back into its output fiber. High-quality mirrors and collimators and precise electrostatic control of the position of each mirror, enable typical switch times of less than 50ms and optical loss that is less than 3.0 dB for CALIENTs complete line of optical circuit switches.